

REPRODUCIBLE RESEARCH PROVENANCE

PETRA ISENBERG

VISUAL ANALYTICS

IN THIS LECTURE

YOU WILL LEARN ABOUT

COMMUNICATING YOUR PROCESS

DETAILS

DIFFICULTIES

HOW TO CONVEY THE ANALYSIS PROCESS?

IN WORDS – TELL IT

PROVIDING DETAIL: CODE, DATA, ...

GENERATE / WRITE REPORTS

WHY CONVEY THE ANALYSIS PROCESS?

SHOW YOUR FINDINGS ARE ROBUST

HIGHLIGHT SUBJECTIVITY

ENABLE IMPROVEMENTS

HELP SOMEONE LEARN ANALYZING

...

PROBLEMS

NOT EASY TO DESCRIBE

PEOPLE MAY NOT UNDERSTAND YOU

LONG ANALYSIS PIPELINES

LOTS OF TRIAL AND ERROR IN ANALYSIS

CONCEPTS

LETS FIRST DISCUSS TWO MAIN CONSIDERATIONS...

REPLICATION VS. REPRODUCIBILITY

REPLICATION


ABILITY OF AN ENTIRE EXPERIMENT / STUDY TO
BE DUPLICATED WITH INDEPENDENT / NEW

DATA

INVESTIGATORS

ANALYSIS METHODS

...



ULTIMATE
STANDARD FOR
STRENGTHENING
SCIENTIFIC
EVIDENCE

REPLICATION WHY?

CHECK IF A FINDING IS ROBUST
IS THIS CLAIM TRUE?

ESPECIALLY IMPORTANT WHEN STUDIES HAVE
BROAD IMPACT
(E.G. ON SOCIETY)

REPLICATION WHEN?

BUT SOMETIMES YOU CAN'T REPLICATE BECAUSE

- YOU DON'T HAVE THE TIME
- OR THE MONEY
- OR THE RESOURCES
- OR THE SITUATION IS UNIQUE

e.g. how would you replicate the Sloan Digital Sky Survey?

IF YOU CAN'T REPLICATE?

WHAT ELSE CAN YOU DO?

LET A STUDY/AN ANALYSIS STAND BY ITSELF?

Do Nothing



Replication

IF YOU CAN'T REPLICATE?

WHAT ELSE CAN YOU DO?

LET A STUDY/AN ANALYSIS STAND BY ITSELF?



REPRODUCIBILITY

REPRODUCIBILITY

ASKS: CAN WE TRUST THIS ANALYSIS?

/SHOULD/ BE MIN STANDARD FOR ANY SCIENTIFIC STUDY

NEW INVESTIGATORS: SAME DATA, SAME METHODS

→ ALLOW FOR VALIDATION OF THE DATA ANALYSIS

WHY?



WHY?

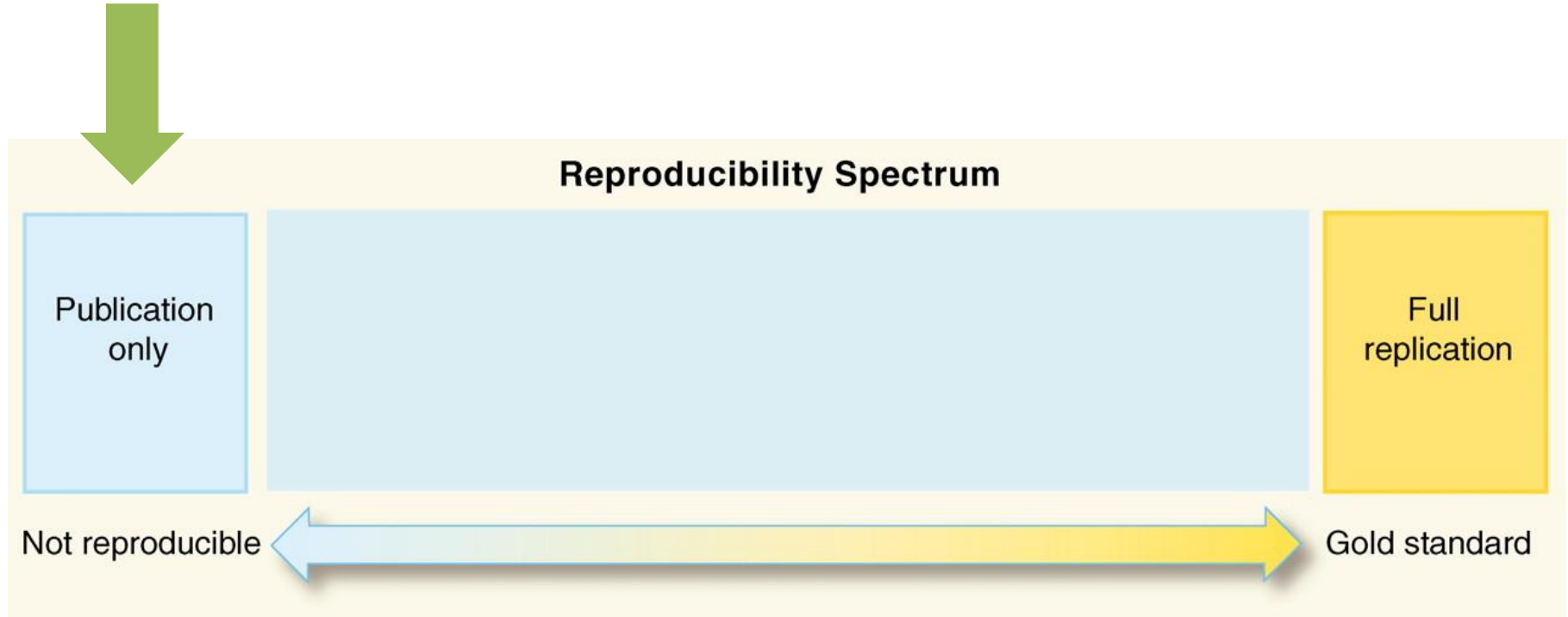
ANOTHER VIDEO FOR YOU TO LOOK AT AT HOME

<https://www.youtube.com/watch?v=eV9dcAGaVU8>

("DECEPTION AT DUKE")

ANALYSIS

(INCL. DATA COLLECTION, CLEANING, ANALYTIC METHODS, FIGURES, ...)



ANALYSIS

(INCL. DATA COLLECTION, CLEANING, ANALYTIC METHODS, FIGURES, ...)



WHAT TO DO?

MAKE YOUR DATA AVAILABLE

MAKE YOUR ANALYSIS METHODS AVAILABLE

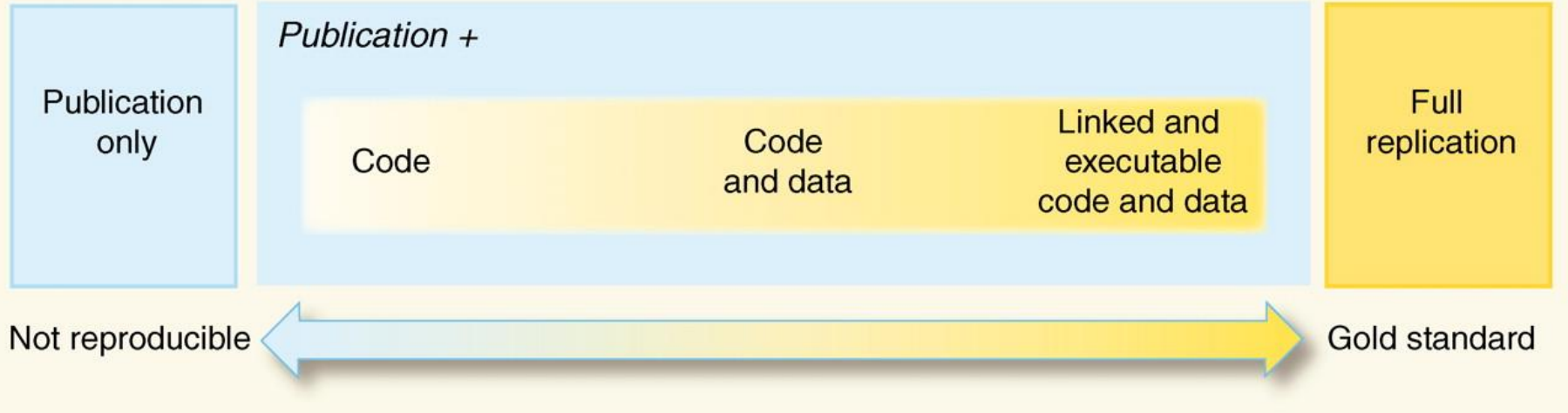
DOCUMENT CODE AND DATA

USE STANDARD MEANS OF DISTRIBUTION

ANALYSIS



Reproducibility Spectrum



WHO IS INVOLVED?

ANALYSTS

WHO WANT TO MAKE THEIR WORK REPRODUCIBLE

READERS

WHO WANT TO REPRODUCE (OR BUILD ON) THE
PREVIOUS ANALYSIS

CHALLENGES

WHAT ARE GOOD TOOLS FOR ANALYSTS?

DOCUMENTATION IS TIME-CONSUMING

NEEDS RESOURCES (WEB SERVERS, ETC.)

WHAT ARE GOOD TOOLS FOR REPRODUCTION?

HOW TO PIECE TOGETHER DATA & CODE

TRYING TO UNDERSTAND WHAT HAPPENED

REPRODUCIBILITY

CONCEPT IMPORTANT TO **ANYONE** CONDUCTING
AN ANALYSIS

BUT: THERE IS NO AGREED-UPON NOTATION FOR
WRITING “INSTRUCTIONS”

REPRODUCIBILITY

For coding environments – like R

BE ORGANIZED

BE ORGANIZED!

YOU WILL DEAL WITH

- DATA (RAW + PROCESSED)
- FIGURES (EXPLORATORY + FINAL)
- CODE (RAW, UNUSED, FINAL, BUGGED, DEBUGGED, ...)
- TEXT (README FILES, ANALYSIS REPORT, DOCUMENTATION)


RAW DATA

SHOULD BE STORED IN YOUR ANALYSIS FOLDER

SHOULD COME WITH README

IF ACCESSED FROM WEB, INCLUDE URL, DESCRIPTION,
AND DATE ACCESSED

PROCESSED DATA
























REMEMBER YOUR
DATA CLEANING
EXERCISES?

SOMETIMES YOU NEED TO TRANSFORM DATA

- NAME PROCESSED DATA TO KNOW WHICH SCRIPT GENERATED IT
- MAKE A README THAT SAYS WHICH SCRIPT/PROCEDURE GENERATED THE FILE
- PROCESSED DATA SHOULD BE READY FOR ANALYSIS







BAD EXAMPLE

| | | | |
|---|--------------------|----------|-----------|
|  coocurrence-author-level1.npy | 3/20/2014 4:33 PM | NPY File | 54,947 KB |
|  coocurrence-author-level1-final clean.npy | 3/20/2014 4:55 PM | NPY File | 53,998 KB |
|  coocurrence-author-level2.npy | 5/7/2014 10:11 AM | NPY File | 191 KB |
|  coocurrence-PCS-all.npy | 5/7/2014 10:11 AM | NPY File | 127 KB |
|  doc-term-level1.npy | 3/20/2014 4:26 PM | NPY File | 21,527 KB |
|  doc-term-level1-final clean.npy | 3/20/2014 4:48 PM | NPY File | 21,341 KB |
|  doc-term-level2.npy | 5/7/2014 10:11 AM | NPY File | 1,267 KB |
|  equivalencematrix.npy | 3/17/2014 1:06 PM | NPY File | 54,039 KB |
|  ieecocoocurrence.npy | 2/11/2014 10:34 AM | NPY File | 29,434 KB |
|  inclusionmatrix.npy | 3/17/2014 1:06 PM | NPY File | 54,039 KB |
|  inspec-controlled-coocurrence.npy | 2/11/2014 1:31 PM | NPY File | 10,369 KB |
|  Matrix5.npy | 3/10/2014 1:24 PM | NPY File | 54,988 KB |
|  Matrix5npy.npy | 3/10/2014 12:46 PM | NPY File | 54,988 KB |
|  Matrix6.npy | 3/10/2014 1:24 PM | NPY File | 54,988 KB |
|  Matrix6npy.npy | 3/10/2014 11:52 AM | NPY File | 54,988 KB |
|  Matrix7.npy | 3/10/2014 1:24 PM | NPY File | 54,988 KB |
|  Matrix7npy.npy | 3/10/2014 11:52 AM | NPY File | 54,988 KB |
|  Matrix8.npy | 3/10/2014 1:24 PM | NPY File | 54,988 KB |
|  Matrix8npy.npy | 3/10/2014 11:52 AM | NPY File | 54,988 KB |
|  Matrix9.npy | 3/10/2014 1:24 PM | NPY File | 54,988 KB |
|  Matrix9npy.npy | 3/10/2014 11:52 AM | NPY File | 54,988 KB |

FIGURES

YOU WILL GENERATE MANY THAT
YOU DON'T NEED

MAKE THE FINAL FIGURES PRETTY,
USE PROPER LABELING AND COLOR,
POSSIBLY CAPTIONS

-  authorkeywordGraph.ai
-  authorkeywordGraph.pdf
-  authorkeywordGraph.svg
-  authorkeywordGraph.svgz
-  authorKeywordGraph2.ai
-  authorKeywordGraph2.pdf

**also name them
properly**

SCRIPTS

CLEARLY COMMENT YOUR FINAL SCRIPTS

WHAT, WHEN, WHY, HOW THROUGHOUT

BIGGER COMMENT BLOCKS FOR WHOLE SECTIONS

INCLUDE PROCESSING DETAILS

CLEAN THE SCRIPT

ONLY INCLUDE CODE FOR FINAL ANALYSIS

GENERAL RECOMMENDATIONS

KEEP TRACK OF WHAT YOU'RE DOING

E.G. USE VERSION CONTROL SYSTEMS

SAVE AS MUCH CODE AS POSSIBLE AS LITTLE OUTPUT
AS NECESSARY

SAVE DATA IN NON-PROPRIETARY FORMATS

PROBLEMS

IT TAKES A LOT OF EFFORT TO MAKE
DATA/RESULTS AVAILABLE

READERS MUST FIND YOUR STUFF AND PIECE IT
TOGETHER

TYPICALLY DATA, CODE, TEXT ARE NOT LINKED

LITERATE PROGRAMMING

LITERATE PROGRAMMING

explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code (Wikipedia)

YOU WRITE CODE TO DO AN ANALYSIS

COMPUTE RESULTS

GENERATE DATA TABLES

...

YOU ALSO WRITE A DOCUMENT – TEXT CHUNKS SURROUNDING YOUR ANALYSIS
CODE

EXPLAIN YOUR ANALYSIS

FORMAT YOUR RESULTS

LITERATE PROGRAMS

USE A DOCUMENTATION LANGUAGE
(HUMAN READABLE)

USE A PROGRAMMING LANGUAGE
(MACHINE READABLE)

HAVE A PRE-PROCESSOR THAT:

WEAVES THE DOC TO PRODUCE HUMAN-READABLE DOCUMENTS (PDF,
HTML, ...)

TANGLES THE DOC TO PRODUCE MACHINE-READABLE DOCUMENTS

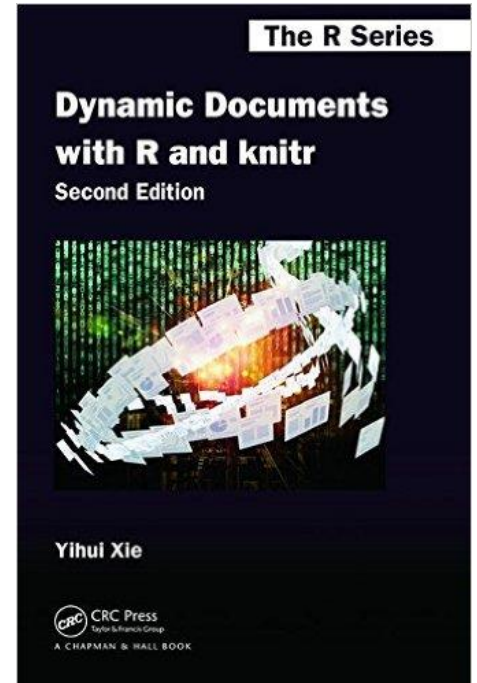
EXAMPLES

FIRST:

WEB (BY DONALD KNUTH, 1981):
PASCAL + TEX

SWEAVE: R + LATEX

KNITR: R + LATEX, MARKDOWN, HTML



```
1 ▾ ---
2 title: "Mayhem at DinoFunWorld"
3 author: "Petra Isenberg"
4 date: "October 5, 2015"
5 output: html_document
6 ▾ ---
7
8 #Merging Data Files with R
9
10 ##Loading Files
11
12 First we will load a file that contains attractions, their ids, and coordinates in the park
13 ▾ ```{r}
14 coordinates <- read.csv("ParkCoordinates.csv")
15 head(coordinates)
16 ^ ```
17
18 Next we will load our data from the data cleaning exercise
19 ▾ ```{r}
20 attractions <- read.csv("AttractionsOCR-txt.csv")
21 head(attractions)
22 ^ ```
23
```

Mayhem at DinoFunWorld

Petra Isenberg

October 5, 2015

Merging Data Files with R

Loading Files

First we will load a file that contains attractions, their ids, and coordinates in the park

```
coordinates <- read.csv("ParkCoordinates.csv")
head(coordinates)
```

```
##           Attraction AttractionID  x  y
## 1 Wrightiraptor Mountain         1 47 11
## 2   Galactosaurus Rage           2 27 15
## 3   Auviolotops Express           3 38 90
## 4           TerrorSaur           4 78 48
## 5   Wendisaurus Chase           5 16 66
## 6   Keimosaurus Big Spin          6 86 44
```

Next we will load our data from the data cleaning exercise

```
attractions <- read.csv("AttractionsOCR-txt.csv")
head(attractions)
```

```
##  AttractionID  ParkArea  Attraction  CategoryNames
## 1            1  Coaster Alley  Wrightiraptor Mountain  Thrill Rides
## 2            2  Coaster Alley   Galactosaurus Rage  Thrill Rides
## 3            3   Tundra Land   Auviolotops Express  Thrill Rides
## 4            4    Wet Land           TerrorSaur  Thrill Rides
## 5            5   Tundra Land   Wendisaurus Chase  Thrill Rides
## 6            6  Coaster Alley   Keimosaurus Big Spin  Thrill Rides
```

+PROS

TEXT AND CODE ALL IN ONE PLACE

ORDER IS MAINTAINED

RESULTS ARE AUTOMATICALLY UPDATED WHEN DATA
CHANGES

CODE NEEDS TO RUN TO PRODUCE THE DOCUMENT

-CONS

DOCUMENTS CAN BECOME DIFFICULT TO READ
WHEN THERE IS A LOT OF CODE

CAN BE SLOW
BUT YOU CAN USE THINGS LIKE CACHING

REPRODUCIBILITY

Is this all we need to understand an analysis and its results?

WHAT ABOUT?

HUMAN PROCESSES SUCH AS

INTERACTIONS WITH GUI SYSTEMS

RESOURCE SHARING/COORDINATION

INSIGHTS AND HYPOTHESES PRODUCED

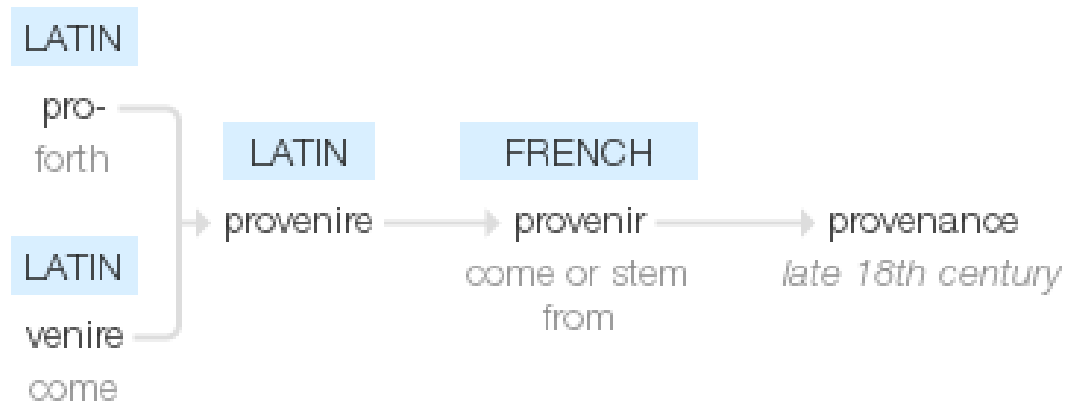
PROVENANCE

A broad concept of “history” in the analysis process

DEFINITION

“ORIGIN, SOURCE”

“THE HISTORY OF OWNERSHIP OF A VALUED OBJECT OR WORK OF ART OF LITERATURE”



PROVENANCE IN VISUAL ANALYTICS

PROVENANCE OF:
DATA
VISUALIZATION
INTERACTIONS
INSIGHTS
RATIONALE

Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes

Eric D. Ragan, Alex Endert, Jibonanda Sanyal, and Jian Chen

Abstract—While the primary goal of visual analytics research is to improve the quality of insights and findings, a substantial amount of research in provenance has focused on the history of changes and advances throughout the analysis process. The term, *provenance*, has been used in a variety of ways to describe different types of records and histories related to visualization. The existing body of provenance research has grown to a point where the consolidation of design knowledge requires cross-referencing a variety of projects and studies spanning multiple domain areas. We present an organizational framework of the different types of provenance information and purposes for why they are desired in the field of visual analytics. Our organization is intended to serve as a framework to help researchers specify types of provenance and coordinate design knowledge across projects. We also discuss the relationships between these factors and the methods used to capture provenance information. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

Index Terms—Provenance, Analytic provenance, Visual analytics, Framework, Visualization, Conceptual model

1 INTRODUCTION

Data visualization and visual analytics combine the power of visualization with advanced data analytics to help people to better understand data and discover meaningful insights. While the goal of visualization research is ultimately to improve the quality of insights and findings, analytic processes are complicated activities involving technology, people, and real world environments. Practical applications encounter problems that extend beyond the integration of any system's analytic models, processing power, visualization designs, and interaction techniques. Visualization systems must also support human processes, which often involve non-standardized methodologies including extended or interrupted periods of analysis, resource sharing and coordination, collaborative work, presentation to different levels of management, and attempts at reproducible analyses [92, 52, 42].

For these reasons, a substantial amount of research in the areas of visualization, data science, and visual analytics has been dedicated to supporting *provenance*, which broadly includes consideration for the history of changes and advances throughout the analysis process (e.g., [34, 73, 37, 21]). It is clear that the research community agrees on the importance of supporting provenance, and many scholars have developed tools and systems that explicitly aim to help analysts record both computational workflows (e.g., [2], 5, 71]) and reasoning processes (e.g., [26, 7]). For example, Vitral tracks steps of the computational workflow during scientific data analysis and visualization, and then provides graphical representations of the workflow through a combination of node diagrams and intermediary visual outputs [5, 14]. Groth and Streefkerk [39] presented another example with a system for recording and annotating stages of view manipulations during a 3D molecule-inspection task. As another example, Del Rio and da Silva [22] designed *Probe-It* to keep track of the data sets that contributed to the creation of map visualizations. Focusing on the provenance of insights, Gotz and Zhou described how the *HARVEST* system records the history of semantic actions during

business and financial analysis activities [37]. These are just a few examples from a large number of visual analytics tools designed to support provenance across a wide range of domains and for different purposes.

As the body of research and existing tools has grown, the community's knowledge of the many factors and goals relevant for effective provenance support has also broadened. However, the variety of perspectives can make it challenging to assess the specific aspects and purposes of provenance that are targeted by any particular project. The term, *provenance*, has been used in a variety of ways to describe different types of origins and histories. For example, the scientific visualization community, especially the simulation and modeling communities, often interpret provenance as the history of computational workflow (e.g., [34]), while other interpretations focus on the history of insights and hypotheses (e.g., [70]). Although many researchers proactively provide clear definitions and explanations of their foci in the provenance research, this does not entirely resolve the challenge of consolidating the variety of interpretations and research outcomes across projects. Different perspectives and applications of concepts become problematic for interpreting and coordinating outcomes from different provenance projects, for communicating ideas within the visualization community, and for allowing new-comers to clearly understand the research space. In our work, we analyzed the different perspectives of provenance that are most relevant to areas of visualization and data analysis.

Our goal in this paper is to organize the different types of provenance information and purposes for why they are desired in information visualization, scientific visualization, and visual analytics. We present an organizational framework as a conceptual model that categorizes and describes the primary components of provenance types and purposes. Further, we discuss the relationships between these factors and considerations when capturing provenance information. Our organizational framework is intended to help researchers specify types of provenance and coordinate design knowledge across projects. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

2 EXISTING PERSPECTIVES OF PROVENANCE

Analytic provenance is a broad and complex concept within the areas of information visualization, data analysis, and data science. In visual data analysis, the concept often includes aspects of the cognitive and interactive processes of discovery and exploration, and also the computational sequences and states traversed to arrive at findings or insights. Prior surveys have presented definitions, categorizations,

- Eric D. Ragan is with Texas A&M University. E-mail: ragan@act.tamu.edu.
- Alex Endert is with Georgia Tech. E-mail: endert@gatech.edu.
- Jibonanda Sanyal is with Oak Ridge National Laboratory. E-mail: sanyal@ornl.gov.
- Jian Chen is with University of Maryland, Baltimore County. E-mail: jchen@umbc.edu.

Manuscript received 11 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 23 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: ircv@computer.org.

PROVENANCE OF DATA

HISTORY OF CHANGES AND MOVEMENT OF DATA
SUBSETTING, MERGING, FORMATTING,...

COUPLED WITH WORKFLOWS
CAPTURES ACTIONS ON DATA

PROVENANCE OF VISUALIZATION

HISTORY OF GRAPHICAL VIEWS AND
VISUALIZATION STATES

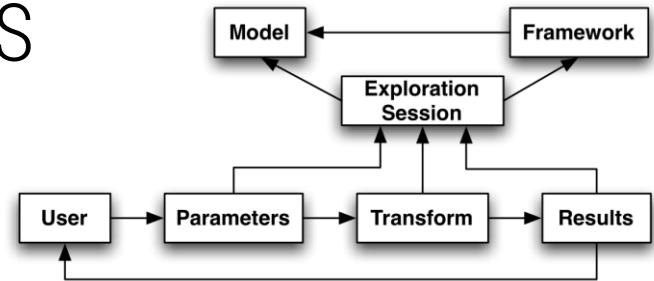
SAVE SCREENSHOTS OR PARAMETERS TO
RECREATE VIEWS/STATES

VISUALIZATION STATES

DESCRIBE VISUALIZATION AS CHAIN
OF VISUAL ENCODING OPERATORS

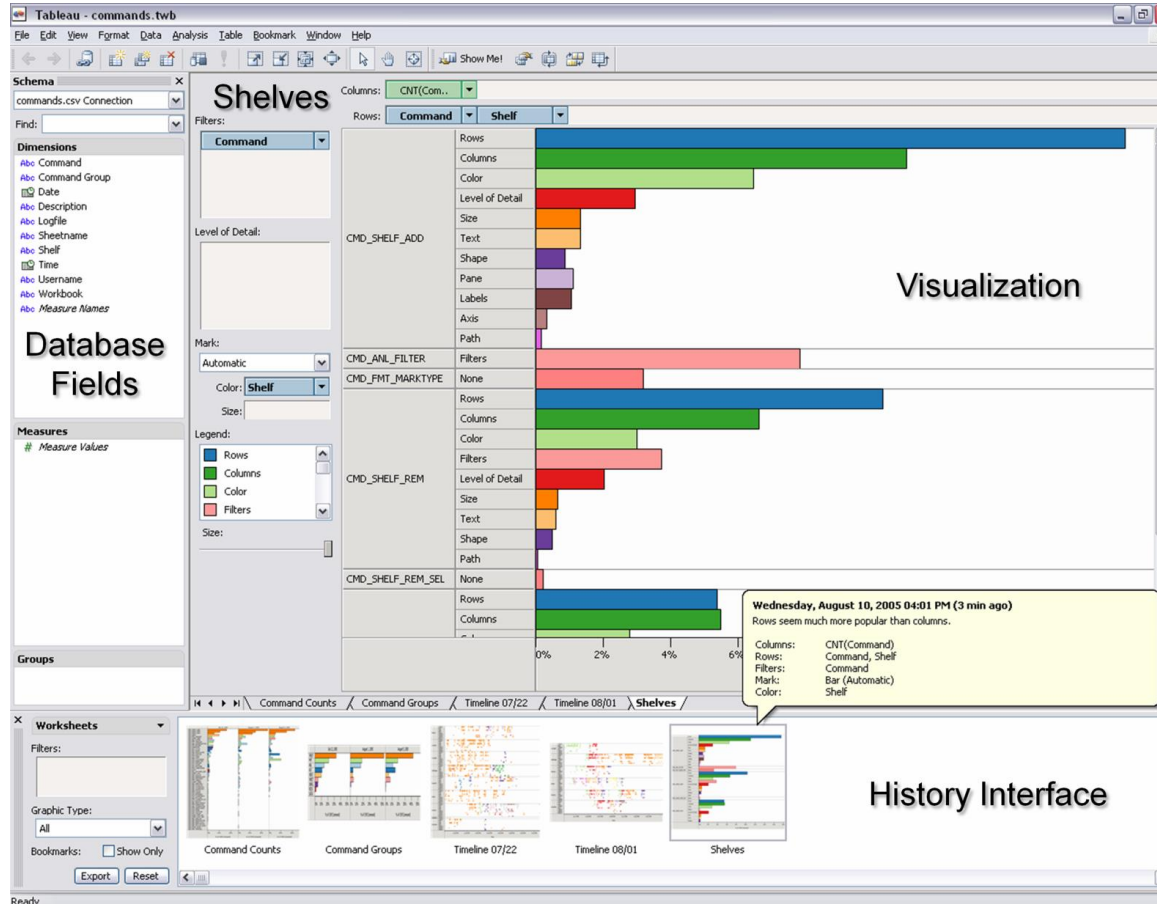
P-SET MODEL:

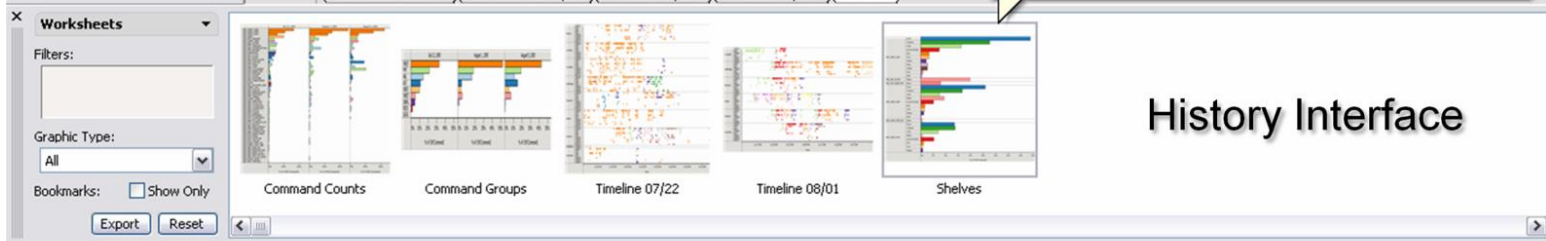
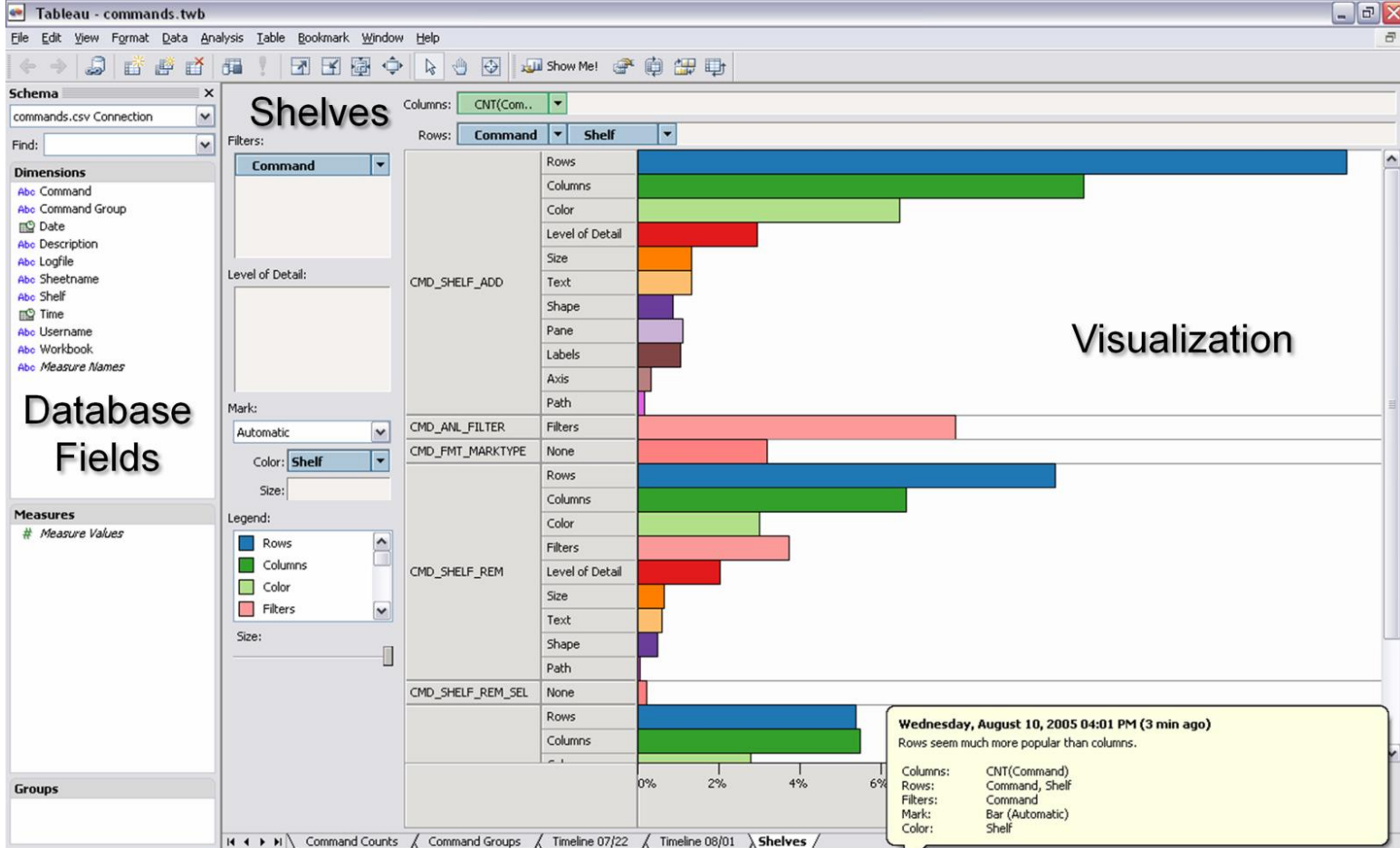
STATE = SET OF PARAMETERS &
ACTIONS AS TRANSFORMATIONS OF
THESE PARAMETERS



A Model and Framework
for Visualization Exploration
T.J. Jankun-Kellym TVCG 2007

VISUALIZATION STATES





Worksheet History ▾

Filters:

Graphic Type:

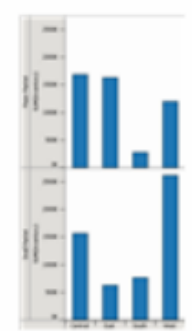
All ▾

Bookmarks: Show Only

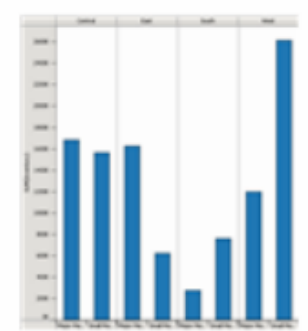
[Export](#) [Reset](#)

| | Major Ma.. | Small Mar.. |
|---------|------------|-------------|
| Central | 1,683,579 | 1,563,045 |
| East | 1,628,963 | 624,021 |
| South | 279,067 | 760,398 |
| West | 1,197,854 | 2,617,410 |

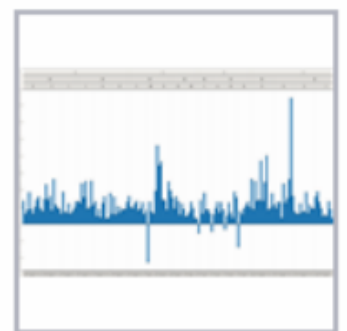
Add Inventory



Show Me!



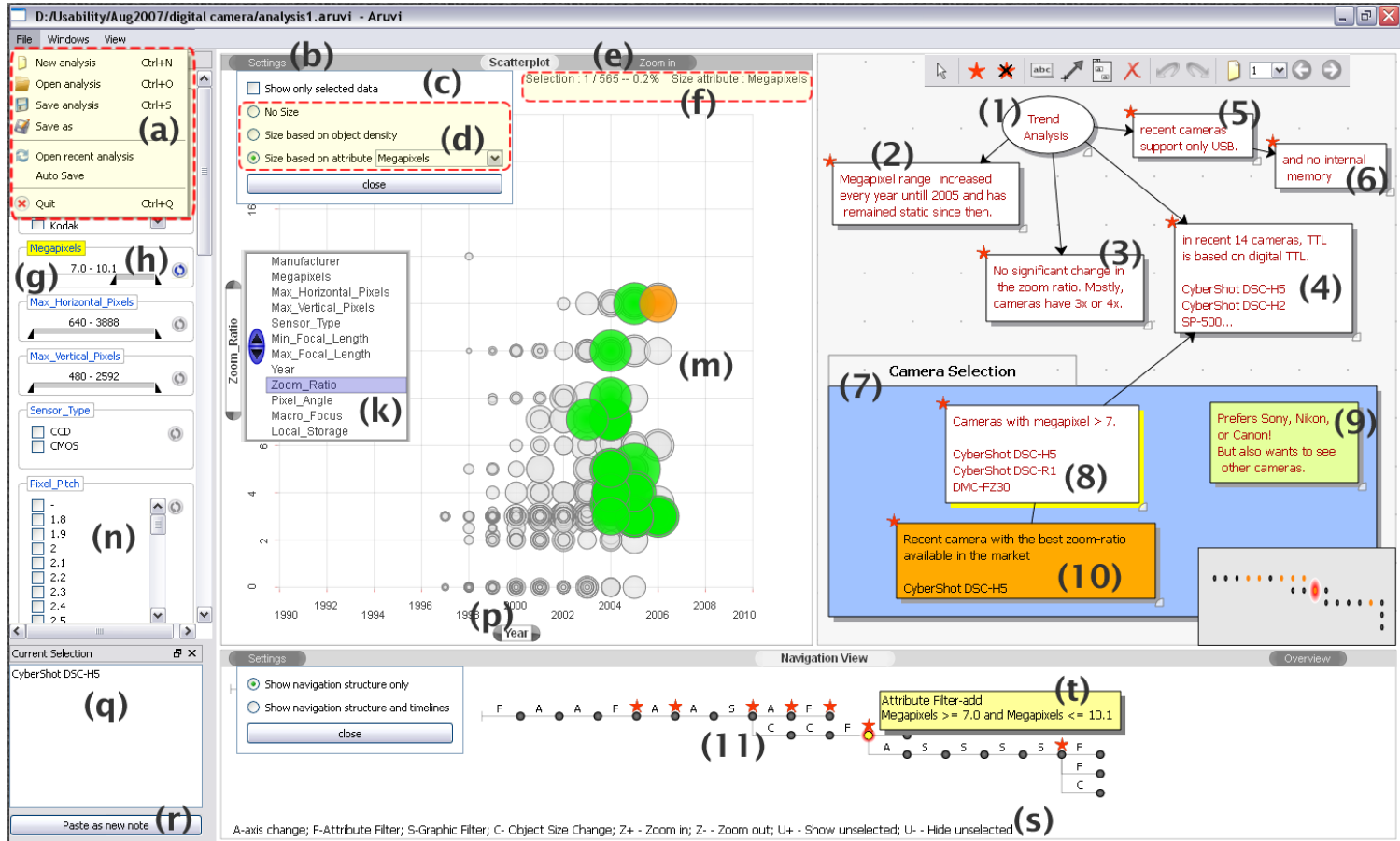
Move Market Size to Columns

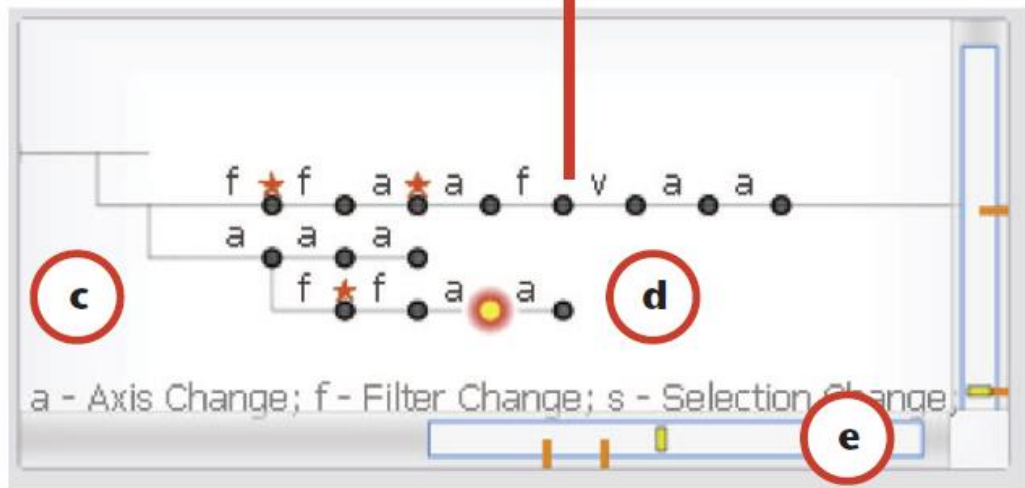
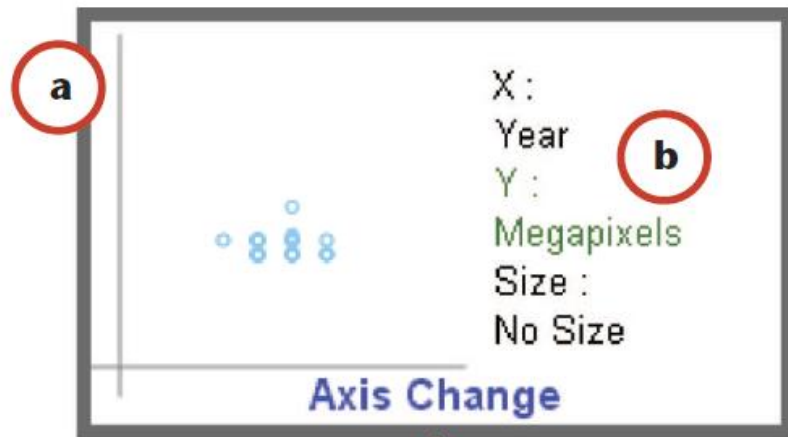


Add Product to Columns



VISUALIZATION STATES





PROVENANCE OF INTERACTIONS WITH A GUI/VIS

HISTORY OF USER INTERACTIONS/COMMANDS

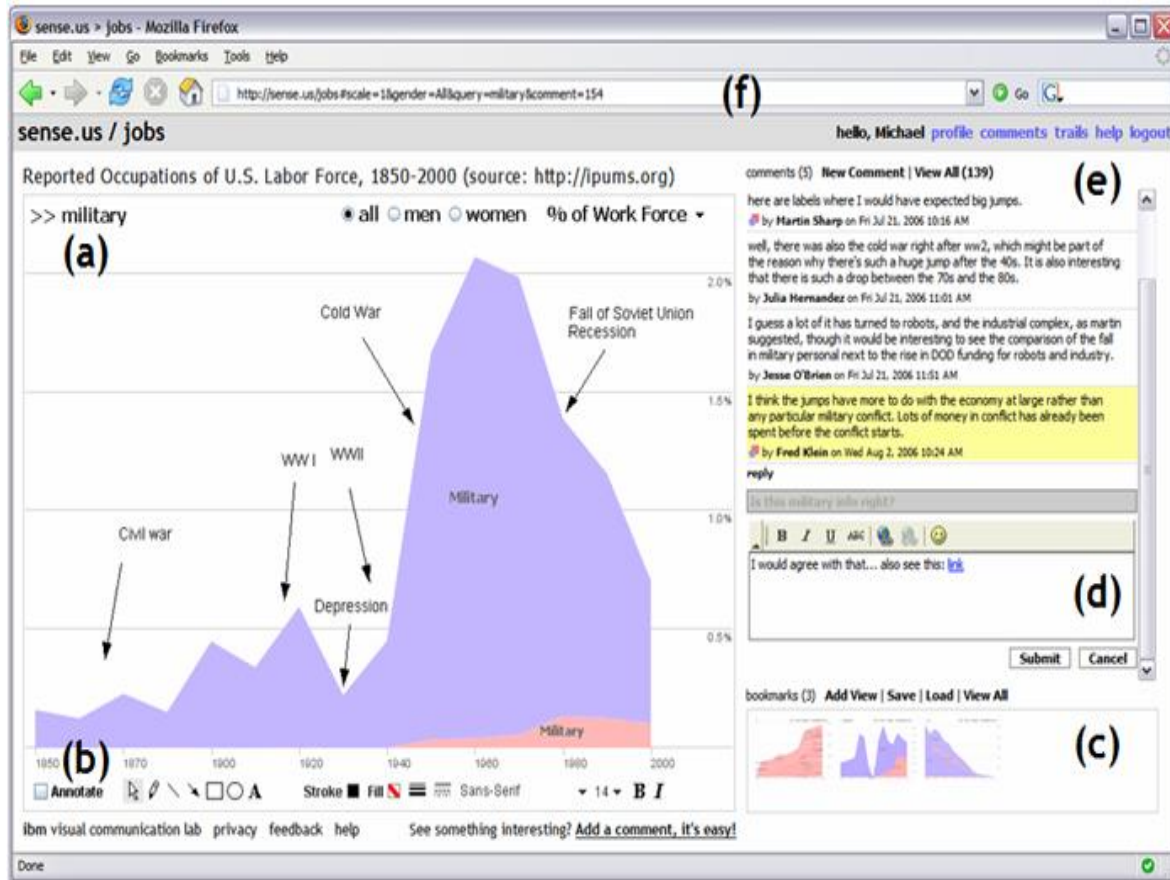
INCLUDES

DATA EXPLORATION INTERACTION (E.G. QUERIES)

ANNOTATION INTERACTIONS

COMMAND HISTORY ACTION (E.G. UNDO/REDO)

(MANUAL) ANNOTATIONS



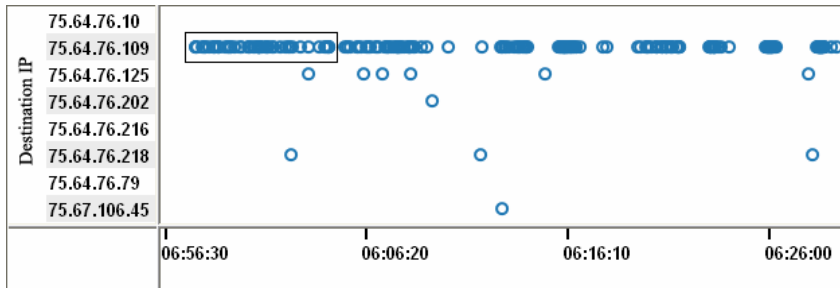
PROVENANCE OF INSIGHT

HISTORY OF COGNITIVE OUTCOMES FROM THE ANALYSIS

DIFFICULT TO CAPTURE, OFTEN MANUALLY ENTERED

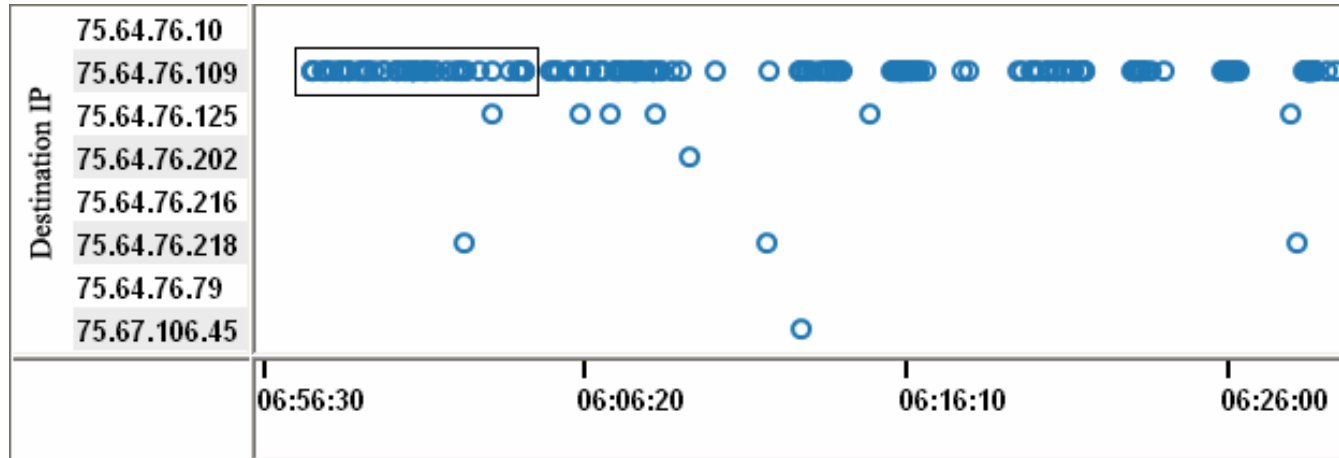
CAPTURING INSIGHT

Network traffic visualization system
Analyst can create logical models of visual discoveries



```
WebCrawl(x1,x2,...) =  
  time_sequence_30s(x1,x2,...) AND  
  more_than_32_events(x1,x2,...) AND  
  identical_source_AS_number(x1,x2,...) AND  
  ( is_web_access_event(x1) AND  
    is_web_access_event(x2) AND ...)
```

CAPTURING INSIGHT



Here: HTTP requests from Google

1) select interesting pattern (burst)

2) system selects a set of predicates (from a list) that are true for these points

CAPTURING INSIGHT

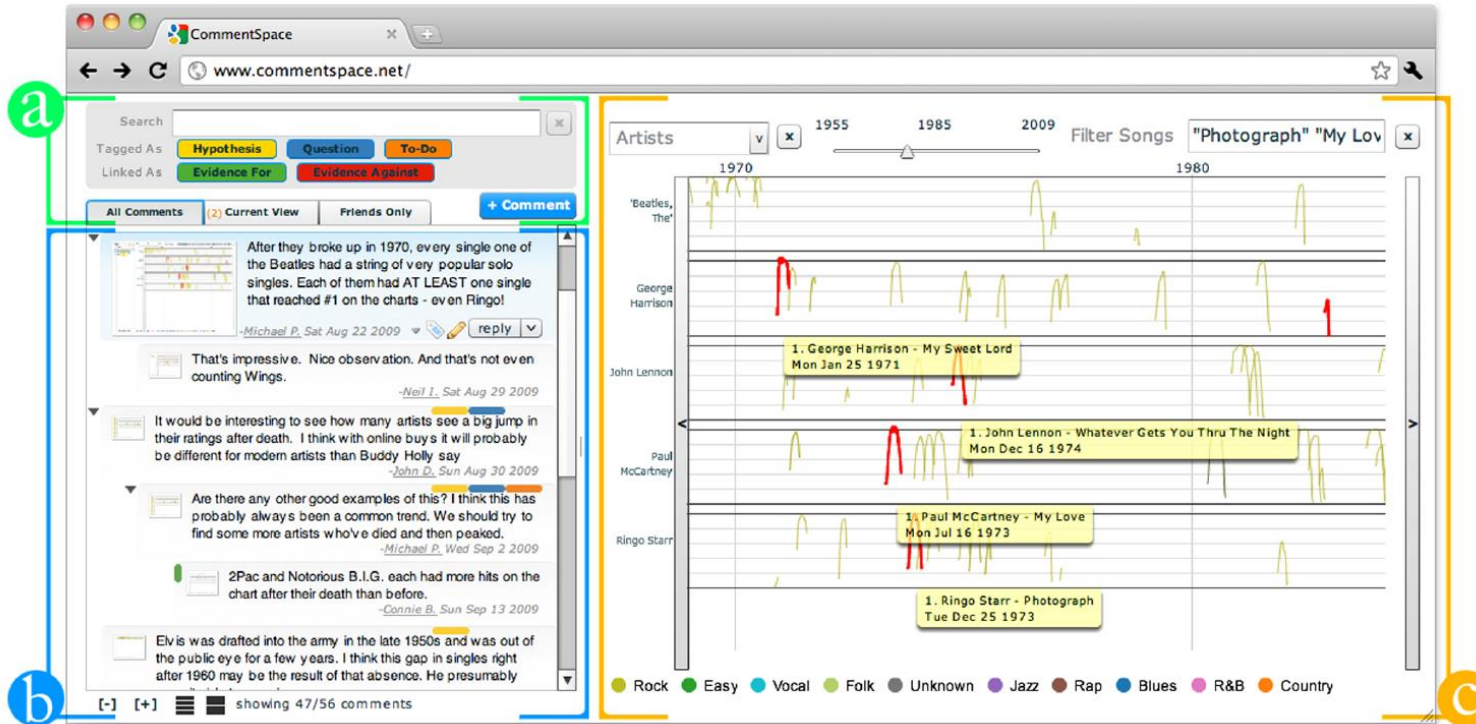
destination_port_80, destination_Stanford,
identical_source_asn, time_sequence_30s,
time_sequence_60s, more_than_4_events,
more_than_32_events

time_sequence_30s(x1,x2,...) AND
more_than_32_events(x1,x2,...) AND
identical_source_AS_number(x1,x2,...) AND
(is_web_access_event(x1) AND
is_web_access_event(x2) AND ...)

selected predicates

analyst modifies list, adds
conjunctions
and looks at visual feedback to
see if pattern is correctly
identified

CAPTURING INSIGHT



PROVENANCE OF RATIONALE

CAPTURE REASONING BEHIND DECISIONS, HYPOTHESES,
INTERACTIONS

GOAL: IDEALLY FIGURE OUT SOMEONE'S ANALYTIC STRATEGY

PROVENANCE IN VISUAL ANALYTICS (RECAP)

PROVENANCE OF:
DATA
VISUALIZATION
INTERACTIONS
INSIGHTS
RATIONALE

Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes

Eric D. Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen

Abstract—While the primary goal of visual analytics research is to improve the quality of insights and findings, a substantial amount of research in provenance has focused on the history of changes and advances throughout the analysis process. The term, *provenance*, has been used in a variety of ways to describe different types of records and histories related to visualization. The existing body of provenance research has grown to a point where the consolidation of design knowledge requires cross-referencing a variety of projects and studies spanning multiple domain areas. We present an organizational framework of the different types of provenance information and purposes for why they are desired in the field of visual analytics. Our organization is intended to serve as a framework to help researchers specify types of provenance and coordinate design knowledge across projects. We also discuss the relationships between these factors and the methods used to capture provenance information. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

Index Terms—Provenance, Analytic provenance, Visual analytics, Framework, Visualization, Conceptual model

1 INTRODUCTION

Data visualization and visual analytics combine the power of visualization with advanced data analytics to help people to better understand data and discover meaningful insights. While the goal of visualization research is ultimately to improve the quality of insights and findings, analytic processes are complicated activities involving technology, people, and real world environments. Practical applications encounter problems that extend beyond the integration of any system's analytic models, processing power, visualization designs, and interaction techniques. Visualization systems must also support human processes, which often involve non-standardized methodologies including extended or interrupted periods of analysis, resource sharing and coordination, collaborative work, presentation to different levels of management, and attempts at reproducible analyses [92, 52, 42].

For these reasons, a substantial amount of research in the areas of visualization, data science, and visual analytics has been dedicated to supporting *provenance*, which broadly includes consideration for the history of changes and advances throughout the analysis process (e.g., [34, 73, 37, 21]). It is clear that the research community agrees on the importance of supporting provenance, and many scholars have developed tools and systems that explicitly aim to help analysts record both computational workflows (e.g., [21, 5, 71]) and reasoning processes (e.g., [26, 71]). For example, Vitral tracks steps of the computational workflow during scientific data analysis and visualization, and then provides graphical representations of the workflow through a combination of node diagrams and intermediary visual outputs [5, 14]. Groth and Steffler [39] presented another example with a system for recording and annotating stages of view manipulations during a 3D molecule-inspection task. As another example, Del Rio and da Silva [22] designed *Probe-It* to keep track of the data sets that contributed to the creation of map visualizations. Focusing on the provenance of insights, Gotz and Zhou described how the *HARVEST* system records the history of semantic actions during

business and financial analysis activities [37]. These are just a few examples from a large number of visual analytics tools designed to support provenance across a wide range of domains and for different purposes.

As the body of research and existing tools has grown, the community's knowledge of the many factors and goals relevant for effective provenance support has also broadened. However, the variety of perspectives can make it challenging to assess the specific aspects and purposes of provenance that are targeted by any particular project. The term, *provenance*, has been used in a variety of ways to describe different types of origins and histories. For example, the scientific visualization community, especially the simulation and modeling communities, often interpret provenance as the history of computational workflow (e.g., [34]), while other interpretations focus on the history of insights and hypotheses (e.g., [70]). Although many researchers proactively provide clear definitions and explanations of their foci in the provenance research, this does not entirely resolve the challenge of consolidating the variety of interpretations and research outcomes across projects. Different perspectives and applications of concepts become problematic for interpreting and coordinating outcomes from different provenance projects, for communicating ideas within the visualization community, and for allowing new-comers to clearly understand the research space. In our work, we analyzed the different perspectives of provenance that are most relevant to areas of visualization and data analysis.

Our goal in this paper is to organize the different types of provenance information and purposes for why they are desired in information visualization, scientific visualization, and visual analytics. We present an organizational framework as a conceptual model that categorizes and describes the primary components of provenance types and purposes. Further, we discuss the relationships between these factors and considerations when capturing provenance information. Our organizational framework is intended to help researchers specify types of provenance and coordinate design knowledge across projects. In addition, our organization can be used to guide the selection of evaluation methodology and the comparison of study outcomes in provenance research.

2 EXISTING PERSPECTIVES OF PROVENANCE

Analytic provenance is a broad and complex concept within the areas of information visualization, data analysis, and data science. In visual data analysis, the concept often includes aspects of the cognitive and interactive processes of discovery and exploration, and also the computational sequences and states traversed to arrive at findings or insights. Prior surveys have presented definitions, categorizations,

- Eric D. Ragan is with Texas A&M University. E-mail: ragan@acm.org.
- Alex Endert is with Georgia Tech. E-mail: endert@gatech.edu.
- Jibonananda Sanyal is with Oak Ridge National Laboratory. E-mail: sanyal@ornl.gov.
- Jian Chen is with University of Maryland, Baltimore County. E-mail: jichen@umbc.edu.

Manuscript received 11 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 23 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: ircv@computer.org.

WHAT TO DO WITH PROVENANCE INFORMATION?

PROVENANCE PURPOSES

RECALL

MEMORY OF STATES OF ANALYSIS

REPRODUCIBILITY

REPRODUCE STEPS/WORKFLOW

ACTION RECOVERY

UNDO/REDO, BRANCHING

PROVENANCE PURPOSES

COLLABORATIVE COMMUNICATION

SHARE INFO WITH OTHERS

PRESENTATION

COMMUNICATE INSIGHT/PROGRESSION

META-ANALYSIS

REVIEW THE ANALYTIC PROCESS

PROVENANCE VS. REPRODUCIBILITY

PROVENANCE VS. REPRODUCIBILITY

GOAL OF GENERAL REPRODUCIBILITY: VALIDATE AN ANALYSIS

- BY SHARING DATA & CODE

HOW CAN WE VALIDATE A VISUAL ANALYSIS?

- BY SHARING INTERACTION LOGS? BY SHARING MANUAL ANALYSIS STEPS? ...
- HOW CAN THIS BE DONE IN A MORE GENERAL WAY ACROSS DIFFERENT GUI-BASED TOOLS?

RESOURCES

- SEE SCIENTIFIC REFERENCES ON SLIDES
- REPRODUCIBLE RESEARCH MOOC
COURSERA.ORG (ROGER PENG)

NEXT UP

AFTER THE BREAK

TUTORIAL – REPRODUCIBLE RESEARCH IN R